Continuity between ancient geochemistry and modern metabolism enabled by non-autocatalytic purine biosynthesis

Joshua E. Goldford[1,2,3,*,#], Harrison B. Smith[2,4,*], Liam M. Longo[2,4,*], Boswell A. Wing[5] and Shawn E. McGlynn[2,4,6,#]

[1]Physics of Living Systems, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[2]Blue Marble Space Institute of Science, Seattle, Washington, USA 98154

[3]Geophysical and Planetary Sciences, California Institute of Technology, Pasadena, CA 91125, USA

[4]Earth-Life Science Institute, Tokyo Institute of Technology, Meguro, Tokyo, Japan 152-8550

[5]Department of Geological Sciences, University of Colorado, Boulder, CO 80309, USA

[6]Biofunctional Catalyst Research Team, RIKEN Center for Sustainable Resource Science, Wako, Saitama, 351-0198, Japan

[*]Co-lead authors

[#]To whom correspondence should be addressed: goldford@mit.edu, mcglynn@elsi.jp

Running Title: Non-autocatalytic purine biosynthesis is required for the emergence of ancient metabolism

Keywords: Ancient metabolism, purine biosynthesis, autocatalysis, metabolic evolution

Abstract

A major unresolved question in the origin of life is whether there exists a continuous path from geochemical precursors to the majority of molecules in the biosphere, due in part to the autocatalytic nature of metabolic networks in modern-day organisms and high rates of extinction throughout Earth's history. Here we simulated the emergence of ancient metabolic networks to identify a feasible path from simple geochemical precursors (e.g., phosphate, sulfide, ammonia, simple carboxylic acids, and metals) to contemporary biochemistry, using only known biochemical reactions and models of primitive coenzymes. We find that purine synthesis constitutes a bottleneck for metabolic expansion, and that non-autocatalytic phosphoryl coupling agents are sufficient to enable expansion from geochemistry to modern metabolic networks. Our model predicts distinct phases of metabolic evolution, characterized by the sequential emergence of key molecules (carboxylic acids, amino acids, sugars), purines/nucleotide cofactors (ATP, NAD^+), flavins, and quinones, respectively. Early phases in the resulting expansion are associated with enzymes that are metal-dependent and structurally symmetric, consistent with models of early biochemical evolution. The production of quinones in the last phase of metabolic expansion permits oxygenic photosynthesis and the production of O_2, leading to a >30% increase in biomolecules. These results reveal a feasible trajectory from simple geochemical precursors to the vast majority of core biochemistry.

Introduction

Metabolism has been proposed to have emerged as a consequence of nascent life deriving material and energy from the surrounding geochemistry (*1*–*3*). However, the transition from geochemistry to biochemistry is poorly understood, due in part to a great uncertainty in the structure of ancient metabolic networks (*4*). In particular, chemical reactions that are unrelated to modern biochemistry have been invoked as missing steps in early biosynthetic pathways (*5*–*7*), suggesting that records of these chemical transformations were lost throughout the history of evolution, through the emergence and sophistication of protein catalysts. Nevertheless, it is unclear to what degree ancient metabolism has been lost, and whether intermediate stages can be excavated from the extant biosphere.

Many lines of evidence suggest that recovering continuity between ancient geochemistry and extant biochemistry might be impossible without the inclusion of a vast number of abiotic chemical reactions unrelated to modern biology. First, a recent analysis of organismal-scale metabolic networks reveals a high prevalence of autocatalytic subnetworks, in which the generation of several key biochemical coenzymes are required for their own synthesis (*8*, *9*). Although autocatalysis has been proposed to have been a necessary feature for early evolutionary processes (*8*, *10*–*12*), the widespread occurrence of such network motifs presents a problem for the initial emergence of ancient metabolism. Second, as the early Earth environment was drastically different from the extant biosphere, high rates of extinction and evolutionary forces like drift could have eroded an early record of ancient biochemistry throughout the course of Earth's history (*13*, *14*). Lastly, several recent studies simulating the emergence of metabolic networks from geochemistry only recovered small networks, on the order of 10% of contemporary biochemistry (*15*, *16*). Taken together, it remains unclear to what extent "extinct" biochemistry is necessary to enable the generation of modern metabolism from early Earth environments.

Beyond addressing the question of continuity, constructing a model for the emergence of biochemical transformations enables us to address key evolutionary questions. For example, resolving the order by which biochemical reactions emerged in time can report on how metabolic pathways evolve more broadly: Models of metabolic pathway evolution have invoked several mechanisms, ranging from sequential models, where reactions emerge in the order they appear in the reaction pathway, to mosaic models, where the order of reaction emergence is uncoupled from the order of reactions in the extant pathway (*17*, *18*). Although studies have shown support for various models of evolution for specific pathways (*18*), a broad, biosphere-scale analysis of the relative occurrence of various modes of metabolic pathway evolution is lacking. Furthermore, the time ordering of biochemical transformations enables generating hypotheses on the relative emergence of metabolic pathways that mediate biogeochemical cycling, such as carbon fixation, which may aid the interpretation of isotopic signatures in the geologic record (*19*).

Here we construct a biosphere-level model of metabolic evolution and uncover a single autocatalytic bottleneck in purine synthesis that prevents the emergence of metabolism from geochemical precursors. We show that including a hypothetical ATP-independent pathway for purine biosynthesis enables the continuous expansion of metabolism from simple starting material, and that the ensuing trajectory of metabolic network evolution is correlated with features typically associated with ancient biochemistry. We use this trajectory to resolve key aspects on the nature of metabolic evolution, with a focus on elucidating the mechanisms and order by which metabolic pathways emerged in the biosphere.

Results

Non-ATP coupled purine production enables metabolic expansion to modern biochemistry

To construct a model of the evolutionary history of metabolism at the biosphere scale, we compiled a database of 12,262 biochemical reactions from the KEGG database (Table S1-5, Methods) (*20*). We added additional detailed organic and inorganic cofactor dependencies for

5,745 reactions from Uniprot, Expasy, PDBE, and EBI (Table S3). These dependencies range from inorganic metal ions (e.g., Fe, Mn) to organic molecules (e.g., flavins, quinones) involved in catalysis (see Methods). Using this network, we performed network expansion (Methods, (*15*, *16*, *21–25*)) starting from a set of "seed compounds". Our seed compounds spanned metals and inorganic material (e.g., Fe, Mn, Zn), CO_2, hydrogen sulfide, molecular hydrogen, orthophosphate, ammonia, and 19 organic substrates producible abiotically starting with pyruvate and glyoxylate (*26–29*) (Table S6 and see supplemental text on succinate semialdehyde). Consistent with previous studies (*15*, *16*), we could generate a network of 433 compounds from diverse pathways in central metabolism, including amino acid biosynthesis and some simple organic coenzymes like PLP (Fig. 1A, red line, Table S7). However, as our biosphere-level network consists of >8000 compounds, this scope only constitutes ~5% of all known biochemicals, leaving the vast majority of molecules unreachable from simple seed compounds. Although the inclusion of primitive thioester energy coupling mechanisms and reductants may have been important during the early stages of biochemical evolution (*15*, *16*), these modifications marginally increased the scope of the expansion (*n*=860) (Fig. 1D). Hypothesizing that our results were biased and limited by the inclusion of only cataloged biochemical reactions, we explored the possibility that unknown biochemical reactions could enable more extensive expansion. To investigate this possibility, we included reactions from a database of hypothetical biochemistry (*30*, *31*), which added 20,183 new reactions to our network and increased the total size by a factor of ~2.7. Repeating the expansion with this expanded reaction set resulted in only a slight increase in scope to 476 compounds (Fig. S1), suggesting that neither currently cataloged nor predicted biochemistry contain transformations required to reach the vast majority of known metabolites.

Notably, phosphoribosyl pyrophosphate (PRPP), a key precursor to metabolite classes like purines, was not in the expansion scope, suggesting that a bottleneck in purine production limits expansion. Indeed, the addition of adenine to the seed set resulted in a network of 4311 compounds, or ~50% network coverage (Fig. 1A, gray line), including all major coenzymes (ATP, NAD, CoA, SAM, flavins, pterins, quinones, and heme, see Table S7). To test whether purines were uniquely essential for the expansion to larger networks, we conducted a “rescue” experiment: for each of the 8294 single compounds from KEGG that were not produced in the original expansion scope (Fig. 1A, red line), we added one to the original seed set, and repeated the expansion. We found that only 216 compounds yielded expansion to >4000 compounds, and all of these compounds either contained a purine moiety or were intermediates in *de novo* purine biosynthesis (Fig. 1C).

De novo purine production in extant biochemistry exhibits an autocatalytic dependence in the production of adenosine triphosphate (ATP) and PRPP, due to multiple steps requiring ATP as a phosphorylating agent (Fig. 1B) (*8*, *9*). This autocatalytic dependence may have been relaxed in primitive metabolism, where primitive phosphorylating agents could have been either pyrophosphate or acyl-phosphate (e.g., acetyl phosphate) (*32–34*). We explored this hypothesis by substituting 10 ATP-coupled reactions involved in purine synthesis with pyrophosphate-coupled variants (see Methods). We repeated the expansion of the original seed set with this modified network and found that expansion led to networks consisting of >4000 molecules with simple phosphoryl donors (Fig. 1A, blue line, Fig. S2, Table S7). This network is biased for compounds in pathways that are frequently observed in microbial genomes (Fig. S3) and, on average, compounds in this network were found in 55% of microbial genomes, compared to just 10% for KEGG compounds not included in the expansion scope (Fig. S3). This result suggests that our model of primitive purine biosynthesis enables a feasible path to the most widely used metabolic pathways in the biosphere. Taken together, this model demonstrates that expansion to most of primary metabolism requires only a small number of additional phosphate-coupled reactions involved in *de novo* purine biosynthesis.

Punctuated metabolic innovation is gated by coenzyme synthesis

Along the primitive purine biosynthesis expansion trajectory (Fig. 1A, blue line), compound addition occurs in bursts, representing periods where large numbers of metabolites were added simultaneously. This observation motivated us to explore the hypothesis that production of key intermediate molecules are bottlenecks throughout the expansion. To test this hypothesis, we removed a single metabolite (and all associated reactions) from the network, performed network expansion on this perturbed network, and then repeated this procedure for each metabolite in the expansion scope (Fig. 2A). Of the 4,240 compounds produced during the expansion, only the removal of 147 led to scopes less than 4000 compounds, indicating that the vast majority of perturbations had little effect on the network size. By plotting the size of the network perturbed by removing a compound (*y*-axis) versus the iteration where the compound was made in the original expansion (*x*-axis), we found that removing certain compounds produced networks of similar size to networks generated before the compounds were produced in the unperturbed expansion (Fig. 2A, red dots). This result suggests that these compounds represent a specific class of "bottleneck" molecules that restrict expansion and are made exactly when they are essential for further expansion of the network (Fig. 2A, red dots). Interestingly, several of these molecules were coenzymes like PLP, ATP, NAD, CoA, and Thiamine diphosphate (ThDP), or intermediates in the biosynthesis of key coenzyme classes like isoprenoids (isopentenyl diphosphate).

As noted in previous studies, the inclusion of various coenzyme classes has a large effect on metabolic network expansion (*21*, *23*), motivating the hypothesis that the observed peaks were induced by the production of key metabolites. To explore this possibility, we simultaneously added 8 key coenzymes (ATP, NAD, CoA, PLP, FAD, SAM, ThDP, and cobalamin) as seed molecules and reran the expansion. Including these coenzymes collapsed the multiple peaks observed in Fig. 1A into just two peaks, where the later peak corresponds to the production of molecular oxygen (Fig. 2B). We further found that addition of these 8 coenzymes to the seed set enabled expansions with fewer iterations relative to randomly chosen sets of 8 compounds with or without oxygen (Monte Carlo permutation test: $P < 10^{-3}$). Altogether, this analysis suggests that the production of major coenzymes are key factors shaping the observed peaks during expansion.

Ancient features of compounds and reactions are correlated with expansion iteration

The last peak shown in Fig. 1A corresponds to the production of molecular oxygen after the emergence of quinones used in modern-day photosystems (Supplemental Text). Since O_2 production was only thought to emerge ~2.4 Gya after the rise of cyanobacteria (*35*, *36*), it is tempting to hypothesize that expansion iteration number may, to a first approximation, represent time. To investigate this possibility, we computed metrics associated with ancient biochemistry and examined how these metrics changed throughout the expansion process.

As suggested recently (*37*), the complexity of molecules (as well as their abundance) might be a useful metric to quantify the degree of evolution of the biosphere. To explore this possibility, we first computed the Bertz complexity of each molecule produced during the expansion with a defined molecular structure (n=3588, Table S8). For each iteration of the expansion, we identified all the compounds produced at that iteration and computed the mean Bertz molecular complexity. The expansion iteration number was positively correlated with mean molecular complexity (Fig. 3A, Pearson's $r = 0.73$, $P < 10^{-17}$) consistent with the hypothesis of increasing molecular complexity during metabolic evolution. In addition to changes in molecular complexity, we found that the average degree of reduction of organic molecules increased throughout the expansion (Fig. S4, Table S9, n=1178, Pearson's $r = 0.89$, $P < 10^{-27}$), potentially a signal of the atmospheric redox state changing throughout Earth's history (*35*, *36*). We also found that the utilization of compounds in prokaryotic genomes decreased throughout the expansion (Fig. 3B, Pearson's $r = -0.81$, $P < 10^{-23}$, see Methods), consistent with the hypothesis that the most ancient parts of metabolism are the most conserved across the tree of life.

Models of ancient metabolic evolution suggest that before the evolution of the modern genetic coding system, "pre-enzymatic" catalysts such as minerals, metal ions, peptides or short RNA polymers served as catalysts for the majority of metabolic reactions (*38*–*41*). It has been proposed that relics of these pre-enzymatic catalysts may still be retained in the enzymatic active sites of many extant enzymes, in the form of metal or iron-sulfur cofactors (*4*, *42*, *43*). As the biosphere became more heavily oxidized, metal solubility decreased (*44*), which increased the selective pressure for enzymes to be less dependent on metal cofactors. Additionally, the sophistication and evolution of enzymes could have alleviated the need for metal or mineral-based catalysis. We thus computed the proportion of reactions at each iteration that were metal cofactor-dependent (metal cofactors listed in Table S10), and observed that this proportion generally decreased as expansion iteration increased (Fig. 3C, Pearson's $r = -0.74$, $P < 10^{-18}$), consistent with the idea that early networks were more dependent on metal-catalyzed reactions.

Beyond molecular complexity and metal-dependence in reaction mechanisms, specific types of protein folds have been hypothesized to have served as ancient catalysts in primitive biochemistry (*45*–*48*). In particular, it has been argued that repetitive/symmetric folds may have been among the first protein folds used in metabolism (*49*), due to their ability to encode complex structures within a short, oligomerizing peptide (*50*–*53*) and their ubiquitous association with diverse metabolic processes (*45*, *54*, *55*). To explore the possibility that reactions appearing early in the expansion are catalyzed by enzymes with greater structural repetition, we computed a structural repetitiveness score for the set of protein folds associated with a reaction (*56*), and the mean repetitiveness score at each expansion iteration (Table S11, see Methods). Consistent with our expectation, we observed a significant decrease in mean repetitiveness throughout the expansion trajectory (Fig. 3D, Pearson's $r = -0.73$, $P < 10^{-17}$). We then quantified the portion of reactions in the network that were dependent on the TIM barrel fold (reaction fold associations listed in Table S12), a protein fold previously suggested to be among the most ancient enzymes (*46*). Similarly to our previous analysis of protein fold repetitiveness, we found that TIM-barrel usage decreased after the generation of ATP (Fig. S5A, Mann-Whitney *U* test: $P < 10^{-3}$). Analysis of nucleotide binding folds (e.g., P-loop NTPases and Rossman folds) show a marked increase after the addition of ATP to the network (Fig. S5B-C, Mann-Whitney *U* tests: $P < 10^{-2}$ and $P < 10^{-7}$, respectively).

Altogether, these analyses show that many features of ancient metabolism – including molecular complexity metrics, metal-cofactor dependencies, and protein fold attributes – are associated with expansion iteration, consistent with the hypothesis that expansion iteration can be viewed as a first-order approximation of time along the trajectory of metabolic evolution.

The modes and order of emergence of extant biochemical pathways

The trajectory of biochemical network evolution permitted us to explore fundamental questions in the history of metabolism, including how metabolic pathways evolve and the specific ordering of distinct metabolic pathways. We first sought to determine the degree to which extant linear metabolic pathways evolved sequentially (either in the forward or reverse direction, see Methods), or non-consecutively, where pathway steps are realized at various times, independent of pathway structure (i.e., the mosaic model). Sequential models of metabolic evolution include the forward direction (also called the Granick model) (*57*), where reactions emerge following the proceeding reaction in the metabolic pathway, and the reverse direction, where reactions are added sequentially in a "retrograde" fashion (*58*). For the 113 linear metabolic pathways in the KEGG database (see Table S13), we computed the iterations where each pathway step was feasible in the expansion and classified these pathways based on the spearman rank correlation between iteration number and when each step was feasible (Methods, Fig. 4A). Surprisingly, we found that the majority (71/113) of these pathways do not emerge sequentially, but rather emerge in a discontinuous fashion, consistent with a mosaic model of metabolic evolution.

We next sought to order the emergence of specific extant metabolic pathways. We constructed logical relationships between reactions and pathways accounting for redundancies and co-dependencies (see Methods), and computed the iteration number where each of 221 metabolic pathways (KEGG modules) were feasible (Table S14). Pathway emergence is distributed throughout the entire expansion (Fig. 4B). Before ATP is generated in the network, only the reductive and non-oxidative pentose phosphate modules, PLP biosynthesis and cysteine biosynthesis are fully enabled (Table S14). This model can be used to generate hypotheses for the ordering of pathways involved in key biogeochemical cycling in the biosphere, such as carbon fixation. Of the seven known carbon fixation pathways, the Calvin cycle was predicted to become feasible first, preceding both the reductive tricarboxylic acid cycle (rTCA) cycle and the reductive acetyl-CoA pathway, which have both been predicted to be operative in the last universal common ancestor (Fig. 4C) (*59–61*).

Discussion

Continuity between early Earth geochemistry and modern metabolism is a necessary condition for any theory of the emergence and evolution of the biosphere on Earth. Whereas demonstrating a continuous trajectory does not rule out possibly critical, yet extinct chemistry (for example (*5*, *6*)), it does relax the requirement to invoke such transformations as essential for biochemical evolution. Furthermore, whereas many other feasible trajectories may enable this observed continuity, the trajectory in Fig. 1A recapitulates several features associated with time and Earth history. Our model suggests that, while species extinction is likely ubiquitous throughout time, extinction of primary metabolic biochemistry may in fact be quite rare. With an approximate time ordering of metabolic exploration, we can explicitly probe the ordering of metabolic pathways and key biochemical innovations in the biosphere. Future work may link biochemical pathway predictions to isotopic fractionation records, although it is unclear how robust these signals are to evolutionary changes in enzyme kinetics over long time-scales (*62*, *63*).

Our model predicts that peaks of metabolic expansion may be characteristic of evolving metabolic networks. Indeed, we found that this punctuated structure is robustly observed when perturbing the initial seed set composition (Fig. S6), or when we include a database of predicted extant yet undiscovered biochemistry (Fig. S1). However, the particular structure and trajectory in Fig. 1A may be sensitive to seed set molecule composition or network structure (Fig. 2B, Fig. S6). In contrast, the emergence of specific metabolic pathways appear to occur with minimal punctuated structure (Fig. 4B), highlighting key differences between the shape of metabolic evolution at the scale of individual reactions and compounds versus pathways. Metabolic pathways also appear to occur as a coalescence of biochemistry emerging at different times throughout the expansion, suggesting the "mosaic" model of metabolic evolution might be the dominant mode of evolution rather than a sequential model. The core metabolic pathways in extant species might have emerged relatively late, with primitive metabolic functions enabled instead by biochemical transformations from disparate metabolic pathways (*16*, *60*).

The model presented in the paper can be improved upon in future work. First, ~50% of the compounds in the KEGG database are still not reachable, suggesting that many reactions are not included in the KEGG database, which may be overcome by incorporating additional biochemical databases (*64–66*) or a much broader scope of hypothetical reactions (Fig. S1) (*67*). Second, a large fraction of biochemical reactions have unknown standard molar free energies, which may be addressable using novel quantum mechanics-based methods for free energy estimation (*68*, *69*). Furthermore, whereas our model explicitly considers dependencies on small organic and inorganic cofactors based on known enzyme mechanisms, future models can incorporate more complex rules encoding dependencies for specific metals or protein-based catalysts.

Although improvements in biochemical annotation and thermodynamic parameter estimation will undoubtedly change details of the trajectory presented in this paper, our analysis suggests that adding an ATP-independent purine biosynthesis pathway to the known catalog of extant biochemistry is a sufficient to enable the continuous trajectory from geochemical precursors to the majority of core metabolism. While the addition of a single hypothetical reaction producing succinate semialdehyde from geochemically available reductants decreases the necessary complexity of the organic seed set to just pyruvate (see Supplemental Information), the inclusion of an ATP-independent route to purine production is essential for seed sets without purines. Compared to the wide occurrence of autocatalytic dependencies at the genome-level (*8*, *9*), there appears to be a paucity of strict autocatalytic dependencies in the biosphere-level metabolism, suggesting that biosphere-level metabolic networks, unlike organismal-scale metabolism, may be the appropriate scale to interrogate evolutionary history of metabolism in deep time. An intriguing possibility is that this single autocatalytic dependency observed here is due to the incomplete characterization of extant biochemical diversity, and that an ancient, non-autocatalytic *de novo* purine synthesis pathway remains hidden in the biosphere.

Acknowledgements

The authors thank Woodward Fisher, Avi Flamholz, and Joan Valentine for valuable discussions. J.E.G. is supported by the Gordon and Betty Moore Foundation as Physics of Living Systems Fellows through grant number GBMF4513. S.E.M. acknowledges support by NSF (Award No. 1724300) “Collaborative Research: Biochemical, Genetic, Metabolic, and Isotopic Constraints on an Ancient Thiobiosphere” and JSPS KAKENHI (Grant Numbers. JP18H01325 and 22H01343). H.B.S acknowledges support from JSPS KAKENHI Grant No. JP19K23459.

Contributions

All authors designed the research. J.E.G, H.B.S. and L.M.L. prepared data. J.E.G. and H.B.S. wrote code and ran simulations. J.E.G. performed analysis. J.E.G., H.B.S. and L.M.L wrote the manuscript. All authors read and approved the final manuscript.

Competing Financial Interests

The authors declare no competing financial interests.

Corresponding Authors

Correspondence to: goldford@mit.edu, mcglynn@elsi.jp

Figures

Figure 1

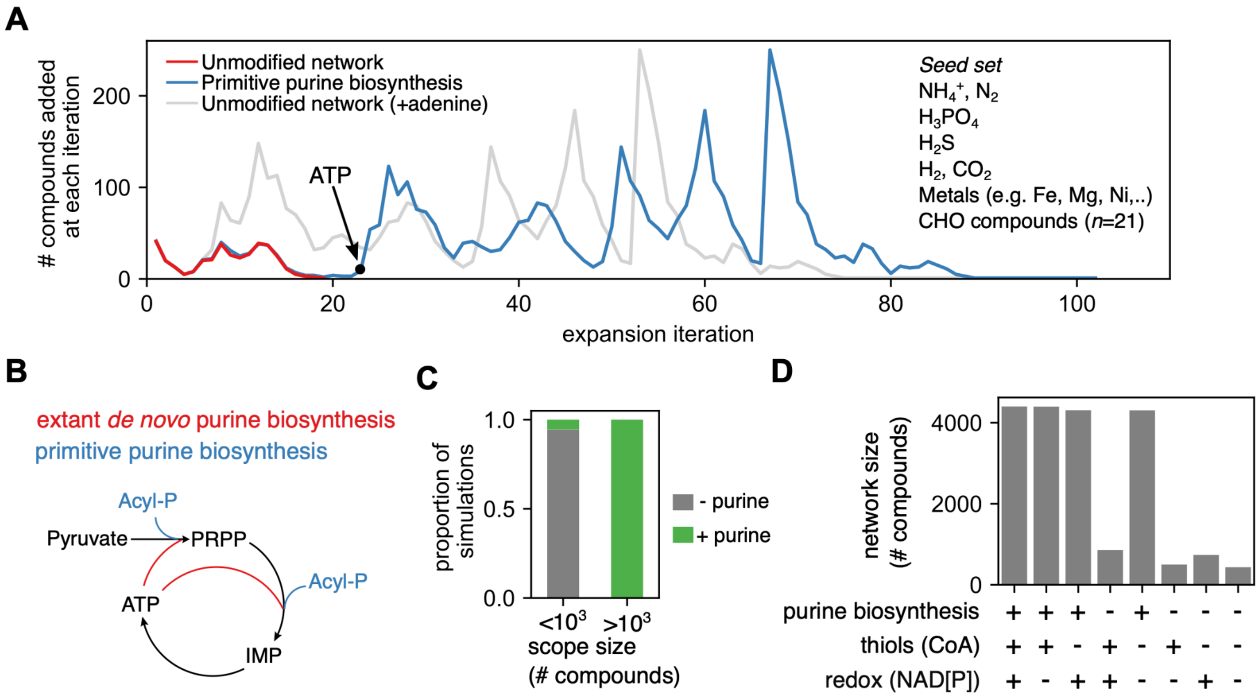

Fig. 1: Primitive purine biosynthesis enables continuity between prebiotic seed molecules and contemporary biochemistry. (A) we performed network expansion with compounds previously hypothesized to have been highly abundant on early Earth (e.g., H_2S, NH_4^+, CO_2, Fe, Mn), as well as molecules producible in simple experiments from pyruvate, glyoxylate, and iron (*26*) (see Table S2 for complete list). For each simulation, we plotted trajectories of the number of compounds added (*y*-axis) at each expansion iteration (*x*-axis), with either an unmodified network without adenine (red line), with adenine (gray line) or a modified network with a model of a primitive acyl-phosphate dependent purine biosynthesis pathway (blue line). (B) The autocatalytic synthesis of ATP in modern metabolism using *de novo* biosynthesis of ATP via IMP (red). We propose that all ATP dependent steps were preceded by acyl-phosphate (Acyl-P) dependent steps (blue, (*32*)). Black lines are steps shared by both the *de novo* and proposed primitive mechanisms. (C) For 8294 compounds, we added each as additional molecules in the

seed set, repeated the expansion, and found that compounds containing purine moieties were required for expansion (see Methods). (D) We repeated expansion with models of primitive purine biosynthesis, primitive thioester coupling or primitive redox systems (in lieu of NAD(P)). While networks increased slightly with primitive redox and thiol coupling, expansion to the >4000 compounds required primitive purine biosynthesis.

Figure 2

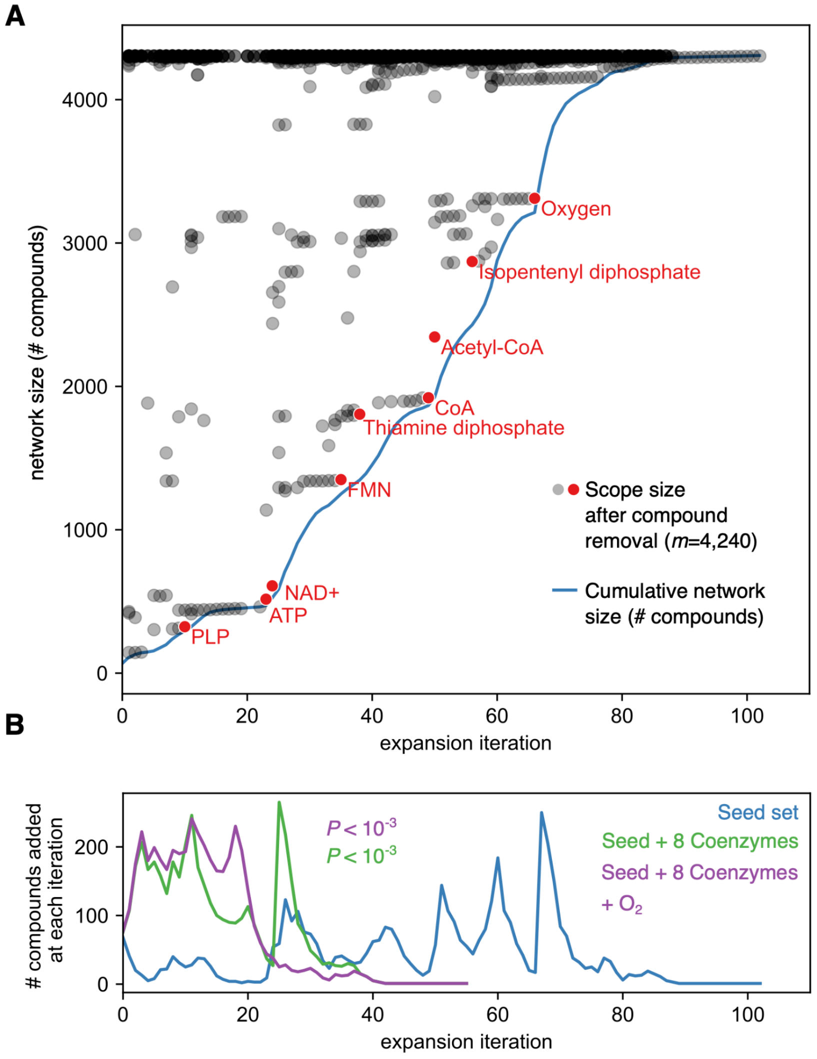

Fig. 2: Peaks during metabolic expansion are associated with coenzyme and oxygen production. (A) For each compound in the expansion scope (*m*=4,240), we removed the

compound and all associated reactions from the network used in Figure 1 and performed expansion. For each compound, we plotted the iteration when the compound is formed in the original expansion (*x*-axis), versus the scope size achieved in the perturbed expansion (*y*-axis). As a reference, we also plotted the cumulative network size (number of compounds) after each iteration for the unperturbed expansion (blue line). Red points are examples of important compounds that become limiting for the expansion close to when they are produced, suggesting they are bottlenecks in the expansion. (B) Trajectories of compounds added at each iteration during network expansion with the base seed set (*n*=70, blue line), base seed set with eight additional coenzymes (ATP, NAD, CoA, PLP, FAD, SAM, ThDP and Cobalamin), (*n*=78, green line), and the base seed set with additional coenzymes and molecular oxygen (*n*=79, purple line). We measured the iteration where 90% of the network was recovered, and found that this value was significantly less than simulations with eight randomly selected compounds as additional seed molecules (Monte carlo permutation test: $P < 10^{-3}$).

Figure 3

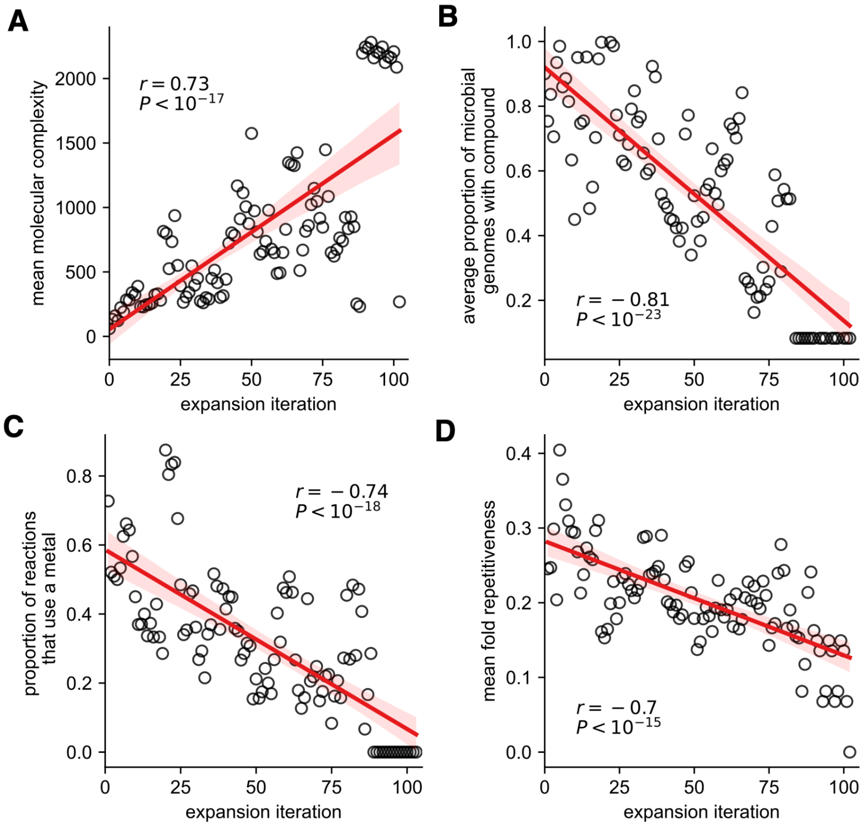

Fig. 3: Ancient biochemical features are correlated with expansion iteration. We computed various biochemical features previously hypothesized to be associated with ancient biochemistry at each iteration during the expansion (*x*-axis, A-D). We computed (A) the average Bertz molecular complexity of molecules produced at each expansion (*y*-axis), (B) the average proportion of microbial genomes using each compound produced during each expansion step (*y*-axis), (C) the proportion of reactions that use an enzyme dependent on a metal cofactor (*y*-axis), and (D) the mean repetitiveness for protein folds associated with each reaction utilized at each expansion iteration (*y*-axis) (see Methods).

Figure 4

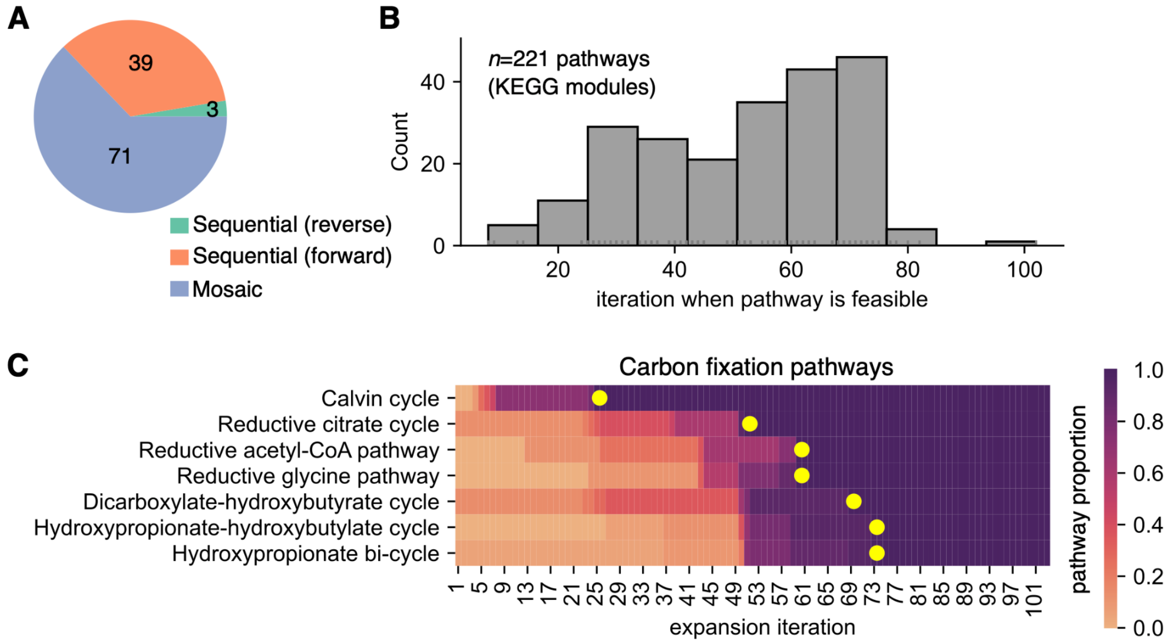

Fig. 4: Timing the emergence of metabolic pathways. (A) For 113 linear metabolic pathways in the KEGG database (Table S13), we classified each pathway as emerging sequentially (either via the forward or reverse direction), or non-sequentially (i.e., Mosaic, blue). We found that 62% (71/113) of these pathways emerged non-sequentially. (B) For 221 KEGG modules, we computed and plotted the distribution of the minimum expansion iteration when each module became feasible (Table S14). (C) A heatmap showing the proportion of feasible steps (color) of seven carbon fixation pathways (*y*-axis) throughout the expansion (*x*-axis). The yellow dots represent the iteration when all steps in the pathway are feasible.

References

1. E. Smith, H. J. Morowitz, *The Origin and Nature of Life On Earth* (Cambridge University Press, Cambridge, United Kingdom, ed. 1st, 2016).

2. E. Smith, H. J. Morowitz, Universality in intermediary metabolism. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 13168–13173 (2004).

3. H. Hartman, Speculations on the origin and evolution of metabolism. *J. Mol. Evol.* **4**, 359–370 (1975).

4. J. E. Goldford, D. Segrè, Modern views of ancient metabolic networks. *Current Opinion in Systems Biology* (2018) (available at https://www.sciencedirect.com/science/article/pii/S2452310017302196).

5. B. H. Patel, C. Percivalle, D. J. Ritson, C. D. Duffy, J. D. Sutherland, Common origins of RNA, protein and lipid precursors in a cyanosulfidic protometabolism. *Nat. Chem.* **7**, 301–307 (2015).

6. J. Xu, V. Chmela, N. J. Green, D. A. Russell, M. J. Janicki, R. W. Góra, R. Szabla, A. D. Bond, J. D. Sutherland, Selective prebiotic formation of RNA pyrimidine and DNA purine nucleosides. *Nature*. **582**, 60–66 (2020).

7. M. Yadav, S. Pulletikurti, J. R. Yerabolu, R. Krishnamurthy, Cyanide as a primordial reductant enables a protometabolic reductive glyoxylate pathway. *Nat. Chem.* **14**, 170–178 (2022).

8. J. C. Xavier, W. Hordijk, S. Kauffman, M. Steel, W. F. Martin, Autocatalytic chemical networks at the origin of metabolism. *Proc. Biol. Sci.* **287**, 20192377 (2020).

9. J. C. Xavier, S. Kauffman, Small-molecule autocatalytic networks are universal metabolic fossils. *Philos. Trans. A Math. Phys. Eng. Sci.* **380**, 20210244 (2022).

10. S. A. Kauffman, *The origins of order : self-organization and selection in evolution* (Oxford University Press, 1993).

11. S. A. Kauffman, Autocatalytic sets of proteins. *J. Theor. Biol.* **119**, 1–24 (1986).

12. A. Blokhuis, D. Lacoste, P. Nghe, Universal motifs and the diversity of autocatalytic systems. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 25230–25236 (2020).

13. C. R. Marshall, Five palaeobiological laws needed to understand the evolution of the living biota. *Nat Ecol Evol*. **1**, 165 (2017).

14. O. L. Petchey, K. J. Gaston, Extinction and the loss of functional diversity. *Proc. Biol. Sci.* **269**, 1721–1727 (2002).

15. J. E. Goldford, H. Hartman, T. F. Smith, D. Segrè, Remnants of an Ancient Metabolism without Phosphate. *Cell*. **168**, 1126–1134.e9 (2017).

16. J. E. Goldford, H. Hartman, R. Marsland 3rd, D. Segrè, Environmental boundary conditions for the

origin of life converge to an organo-sulfur metabolism. *Nat Ecol Evol*. **3**, 1715–1724 (2019).

17. K. B. Muchowska, S. J. Varma, J. Moran, Nonenzymatic metabolic reactions and life's origins. *Chem. Rev.* (2020) (available at https://pubs.acs.org/doi/abs/10.1021/acs.chemrev.0c00191).

18. H. A. Maeda, A. R. Fernie, Evolutionary History of Plant Metabolism. *Annu. Rev. Plant Biol.* **72**, 185–216 (2021).

19. A. K. Garcia, C. M. Cavanaugh, B. Kacar, The curious consistency of carbon biosignatures over billions of years of Earth-life coevolution. *ISME J.* **15**, 2183–2194 (2021).

20. M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

21. O. Ebenhöh, T. Handorf, R. Heinrich, Structural analysis of expanding metabolic networks. *Genome Inform.* **15**, 35–45 (2004).

22. T. Handorf, O. Ebenhöh, R. Heinrich, Expanding metabolic networks: scopes of compounds, robustness, and evolution. *J. Mol. Evol.* **61**, 498–512 (2005).

23. J. Raymond, D. Segrè, The effect of oxygen on biochemical networks and the evolution of complex life. *Science*. **311**, 1764–1767 (2006).

24. H. Kim, H. B. Smith, C. Mathis, J. Raymond, S. I. Walker, Universal scaling across biochemical networks on Earth. *Sci Adv*. **5**, eaau0149 (2019).

25. T. Tian, X.-Y. Chu, Y. Yang, X. Zhang, Y.-M. Liu, J. Gao, B.-G. Ma, H.-Y. Zhang, Phosphates as Energy Sources to Expand Metabolic Networks. *Life*. **9** (2019), doi:10.3390/life9020043.

26. K. B. Muchowska, S. J. Varma, J. Moran, Synthesis and breakdown of universal metabolic precursors promoted by iron. *Nature*. **569**, 104–107 (2019).

27. S. J. Varma, K. B. Muchowska, P. Chatelain, J. Moran, Native iron reduces CO2 to intermediates and end-products of the acetyl-CoA pathway. *Nature Ecology & Evolution*. **2** (2018), doi:10.1038/s41559-018-0542-2.

28. G. Springsteen, J. R. Yerabolu, J. Nelson, C. J. Rhea, R. Krishnamurthy, Linked cycles of oxidative decarboxylation of glyoxylate as protometabolic analogs of the citric acid cycle. *Nat. Commun.* **9**, 91 (2018).

29. K. B. Muchowska, S. J. Varma, E. Chevallot-Beroux, L. Lethuillier-Karl, G. Li, J. Moran, Metals promote sequences of the reverse Krebs cycle. *Nature Ecology and Evolution*. **1**, 1716–1721 (2017).

30. N. Hadadi, J. Hafner, A. Shajkofci, A. Zisaki, V. Hatzimanikatis, ATLAS of Biochemistry: A Repository of All Possible Biochemical Reactions for Synthetic Biology and Metabolic Engineering Studies. *ACS Synth. Biol.* **5**, 1155–1166 (2016).

31. J. Hafner, H. MohammadiPeyhani, A. Sveshnikova, A. Scheidegger, V. Hatzimanikatis, Updated ATLAS of Biochemistry with New Metabolites and Improved Enzyme Prediction Power. *ACS Synth. Biol.* **9**, 1479–1482 (2020).

32. A. Whicher, E. Camprubi, S. Pinna, B. Herschy, N. Lane, Acetyl Phosphate as a Primordial Energy Currency at the Origin of Life. *Orig. Life Evol. Biosph.* **48**, 159–179 (2018).

33. S. A. Harrison, N. Lane, Life as a guide to prebiotic nucleotide synthesis. *Nat. Commun.* **9**, 5176 (2018).

34. S. Pinna, C. Kunz, A. Halpern, S. A. Harrison, S. F. Jordan, J. Ward, F. Werner, N. Lane, A prebiotic basis for ATP as the universal energy currency. *PLoS Biol.* **20**, e3001437 (2022).

35. D. C. Catling, K. J. Zahnle, The Archean atmosphere. *Sci Adv*. **6**, eaax1420 (2020).

36. W. W. Fischer, J. Hemp, J. E. Johnson, Evolution of oxygenic photosynthesis. *Annual Review of Earth and* (2016) (available at http://www.gps.caltech.edu/~wfischer/pubs/Fischeretal2016a.pdf).

37. S. M. Marshall, C. Mathis, E. Carrick, G. Keenan, G. J. T. Cooper, H. Graham, M. Craven, P. S. Gromski, D. G. Moore, S. I. Walker, L. Cronin, Identifying molecules as biosignatures with assembly theory and mass spectrometry. *Nat. Commun.* **12**, 3033 (2021).

38. J. C. Fontecilla-Camps, Geochemical Continuity and Catalyst/Cofactor Replacement in the Emergence and Evolution of Life. *Angew. Chem. Int. Ed Engl.* **58**, 42–48 (2019).

39. E. J. Milner-White, M. J. Russell, Functional capabilities of the earliest peptides and the emergence of life. *Genes* . **2**, 671–688 (2011).

40. A. D. Goldman, B. Kacar, Cofactors are Remnants of Life's Origin and Early Evolution. *J. Mol. Evol.* (2021), doi:10.1007/s00239-020-09988-4.

41. L. M. Longo, D. Despotović, O. Weil-Ktorza, M. J. Walker, J. Jabłońska, Y. Fridmann-Sirkis, G. Varani, N. Metanis, D. S. Tawfik, Primordial emergence of a nucleic acid-binding protein via phase separation and statistical ornithine-to-arginine conversion. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 15731–15739 (2020).

42. W. Martin, M. J. Russell, On the origin of biochemistry at an alkaline hydrothermal vent. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **362**, 1887–1926 (2007).

43. W. Nitschke, S. E. McGlynn, E. J. Milner-White, M. J. Russell, On the antiquity of metalloenzymes and their substrates in bioenergetics. *Biochim. Biophys. Acta*. **1827**, 871–881 (2013).

44. M. A. Saito, D. M. Sigman, F. M. M. Morel, The bioinorganic chemistry of the ancient ocean: the co-evolution of cyanobacterial metal requirements and biogeochemical cycles at the Archean–Proterozoic boundary? *Inorganica Chim. Acta*. **356**, 308–318 (2003).

45. B.-G. Ma, L. Chen, H.-F. Ji, Z.-H. Chen, F.-R. Yang, L. Wang, G. Qu, Y.-Y. Jiang, C. Ji, H.-Y. Zhang, Characters of very ancient proteins. *Biochem. Biophys. Res. Commun.* **366**, 607–611 (2008).

46. A. D. Goldman, J. T. Beatty, L. F. Landweber, The TIM Barrel Architecture Facilitated the Early Evolution of Protein-Mediated Metabolism. *J. Mol. Evol.* **82**, 17–26 (2016).

47. N. Nath, J. B. O. Mitchell, G. Caetano-Anollés, The Natural History of Biocatalytic Mechanisms. *PLoS Comput. Biol.* **10** (2014), doi:10.1371/journal.pcbi.1003642.

48. M. F. Aziz, K. Caetano-Anollés, G. Caetano-Anollés, The early history and emergence of molecular functions and modular scale-free network behavior. *Sci. Rep.* **6**, 25058 (2016).

49. R. V. Eck, M. O. Dayhoff, Evolution of the structure of ferredoxin based on living relics of primitive amino Acid sequences. *Science*. **152**, 363–366 (1966).

50. J. Lee, M. Blaber, Experimental support for the evolution of symmetric protein architecture from a simple peptide motif. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 126–130 (2011).

51. R. G. Smock, I. Yadid, O. Dym, J. Clarke, D. S. Tawfik, De Novo Evolutionary Emergence of a Symmetrical Protein Is Shaped by Folding Constraints. *Cell*. **164**, 476–486 (2016).

52. S. Yagi, A. K. Padhi, J. Vucinic, S. Barbe, T. Schiex, R. Nakagawa, D. Simoncini, K. Y. J. Zhang, S. Tagami, Seven Amino Acid Types Suffice to Create the Core Fold of RNA Polymerase. *J. Am. Chem. Soc.* **143**, 15998–16006 (2021).

53. M. Seal, O. Weil-Ktorza, D. Despotović, D. S. Tawfik, Y. Levy, N. Metanis, L. M. Longo, D. Goldfarb, Peptide-RNA Coacervates as a Cradle for the Evolution of Folded Domains. *J. Am. Chem. Soc.* **144**, 14150–14160 (2022).

54. R. K. Wierenga, The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Lett.* **492**, 193–198 (2001).

55. L. M. Longo, J. Jabłońska, P. Vyas, M. Kanade, R. Kolodny, N. Ben-Tal, D. S. Tawfik, On the emergence of P-Loop NTPase and Rossmann enzymes from a Beta-Alpha-Beta ancestral fragment. *Elife*. **9** (2020), doi:10.7554/eLife.64415.

56. S. E. Bliven, A. Lafita, P. W. Rose, G. Capitani, A. Prlić, P. E. Bourne, Analyzing the symmetrical arrangement of structural repeats in proteins with CE-Symm. *PLoS Comput. Biol.* **15**, e1006842 (2019).

57. S. Granick, "Evolution of Heme and Chlorophyll" in *Evolving Genes and Proteins*, V. Bryson, H. J. Vogel, Eds. (Academic Press, 1965), pp. 67–88.

58. N. H. Horowitz, On the Evolution of Biochemical Syntheses. *Proc. Natl. Acad. Sci. U. S. A.* **31**, 153–157 (1945).

59. G. Fuchs, Alternative pathways of carbon dioxide fixation: insights into the early evolution of life? *Annu. Rev. Microbiol.* **65**, 631–658 (2011).

60. R. Braakman, E. Smith, The emergence and early evolution of biological carbon-fixation. *PLoS Comput. Biol.* **8** (2012), doi:10.1371/journal.pcbi.1002455.

61. M. C. Weiss, F. L. Sousa, N. Mrnjavac, S. Neukirchen, M. Roettger, S. Nelson-Sathi, W. F. Martin, The physiology and habitat of the last universal common ancestor. *Nature Microbiology*. **1**, 16116 (2016).

62. M. Kędzior, A. K. Garcia, M. Li, A. Taton, Z. R. Adam, J. N. Young, B. Kaçar, Resurrected Rubisco suggests uniform carbon isotope signatures over geologic time. *Cell Rep.* **39**, 110726 (2022).

63. R. Z. Wang, R. J. Nichols, A. K. Liu, A. I. Flamholz, D. M. Banda, D. F. Savage, J. M. Eiler, P. M.

Shih, W. W. Fischer, Evolution of Carbon Isotope Fractionation in Cyanobacteria. *bioRxiv* (2022), p. 2022.06.22.497258.

64. R. Caspi, R. Billington, I. M. Keseler, A. Kothari, M. Krummenacker, P. E. Midford, W. K. Ong, S. Paley, P. Subhraveti, P. D. Karp, The MetaCyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic Acids Res.* **48**, D445–D453 (2020).

65. P. Bansal, A. Morgat, K. B. Axelsen, V. Muthukrishnan, E. Coudert, L. Aimo, N. Hyka-Nouspikel, E. Gasteiger, A. Kerhornou, T. B. Neto, M. Pozzato, M.-C. Blatter, A. Ignatchenko, N. Redaschi, A. Bridge, Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Res.* **50**, D693–D700 (2022).

66. S. M. D. Seaver, F. Liu, Q. Zhang, J. Jeffryes, J. P. Faria, J. N. Edirisinghe, M. Mundy, N. Chia, E. Noor, M. E. Beber, A. A. Best, M. DeJongh, J. A. Kimbrel, P. D'haeseleer, S. R. McCorkle, J. R. Bolton, E. Pearson, S. Canon, E. M. Wood-Charlson, R. W. Cottingham, A. P. Arkin, C. S. Henry, The ModelSEED Biochemistry Database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes. *Nucleic Acids Res.* **49**, D575–D588 (2021).

67. H. MohammadiPeyhani, J. Hafner, A. Sveshnikova, V. Viterbo, V. Hatzimanikatis, Expanding biochemical knowledge and illuminating metabolic dark matter with ATLASx. *Nat. Commun.* **13**, 1560 (2022).

68. A. Jinich, B. Sanchez-Lengeling, H. Ren, A mixed quantum chemistry/machine learning approach for the fast and accurate prediction of biochemical redox potentials and its large-scale application to 315 000 …. *ACS central* (2019) (available at https://pubs.acs.org/doi/abs/10.1021/acscentsci.9b00297).

69. A. Jinich, B. Sanchez-Lengeling, H. Ren, J. E. Goldford, E. Noor, J. N. Sanders, D. Segrè, A. Aspuru-Guzik, A thermodynamic atlas of carbon redox chemical space. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 32910–32918 (2020).

70. E. Noor, H. S. Haraldsdóttir, R. Milo, R. M. T. Fleming, Consistent estimation of Gibbs energy using component contributions. *PLoS Comput. Biol.* **9**, e1003098 (2013).

71. M. E. Beber, M. G. Gollub, D. Mozaffari, K. M. Shebek, A. I. Flamholz, R. Milo, E. Noor, eQuilibrator 3.0: a database solution for thermodynamic constant estimation. *Nucleic Acids Res.* (2021), doi:10.1093/nar/gkab1106.

72. S. H. Bertz, The first general index of molecular complexity. *J. Am. Chem. Soc.* **103**, 3599–3601 (1981).

73. L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. **28**, 3150–3152 (2012).

74. H. Cheng, R. D. Schaeffer, Y. Liao, L. N. Kinch, J. Pei, S. Shi, B.-H. Kim, N. V. Grishin, ECOD: an evolutionary classification of protein domains. *PLoS Comput. Biol.* **10**, e1003926 (2014).

75. H. Cheng, Y. Liao, R. D. Schaeffer, N. V. Grishin, Manual classification strategies in the ECOD database. *Proteins*. **83**, 1238–1251 (2015).

76. J. Mistry, R. D. Finn, S. R. Eddy, A. Bateman, M. Punta, Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121 (2013).

Materials and Methods

Software availability

Code and data are available on the following github repository:

https://github.com/jgoldford/metabolic-continuity

Metabolic network reconstruction

Briefly, we downloaded the KEGG network on May 31, 2021. However, unlike previous studies, we identified *n*=230 reactions that were elementally inconsistent (i.e., the reactants and products did not share the same elements), so these were removed. Reactions that were stoichiometrically imbalanced, or relied on metabolites with unknown "R-groups", were kept.

Missing reaction and cofactor annotation

Cofactors and EC-cofactor dependencies were identified from Expasy (via the `cofactorLabel` field and comments), PDBE (via the cofactor summary), Uniprot (via cofactor/EC numbers), and KEGG (via comments and reaction definitions). For more details, see the SI section titled "Cofactor Annotations". We added several missing reactions in KEGG to ensure production of key, generic metabolites, such as quinones, ferricytochrome, etc. We describe each modification in the SI section titled "Addition of missing reactions."

Thermodynamic constraints

Standard molar free energies were estimated using the component contribution method (*70*) and the eQuilibrator python API (*71*) using the following parameters: pH=7.0, pMg=3.0, ionic strength of 0.25M and *T*=298.15. For each reaction with a free energy estimate at standard molar conditions, we estimated the maximum and minimum reaction free energy for both the forward and reverse reactions. We computed these values by assuming that compound concentration ranges were between 10^{-7} and 10^{-1} M, and computed the maximum and minimum reaction free energies as described previously (*16*). For all calculations, we assumed activity coefficients were well approximated by concentrations. All forward or reverse reactions with minimum free energies above zero were removed from the network. For KEGG reactions with no free energy estimate (*n*=839), we kept both the forward and reverse reaction in the network.

Modeling primitive purine biosynthesis reactions

We modeled the primitive purine biosynthesis by adding 10 new reactions to the model. We replaced all ATP (GTP) with Pyrophosphate (PP_i) (C00013) and ADP (GDP) with orthophosphate (C00009) and recomputed standard molar free energies using eQuilibrator for the following KEGG reactions: ATP:D-ribulose-5-phosphate 1-phosphotransferase (R01523, EC 2.7.1.19), 5-Phospho-D-ribosylamine:glycine ligase (R04144, EC 6.3.4.13), ATP:pyruvate 2-O-phosphotransferase (R00200, EC 2.7.1.40), Ribose-5-phosphate:ammonia ligase (R01053, EC 6.3.4.7), formate:5-amino-1-(5-phospho-D-ribosyl)imidazole-4-carboxamide ligase (R06975, EC 6.3.4.23), ATP:ribose-1,5-bisphosphate phosphotransferase (R06836, EC 2.7.4.23), 1-(5-Phosphoribosyl)-5-amino-4-carboxyimidazole:L-aspartate ligase (R04591, EC 6.3.2.6), 5'-Phosphoribosylformylglycinamide:L-glutamine amido-ligase (R04463, EC 6.3.5.3), formate:N1-(5-phospho-beta-D-ribosyl)glycinamide ligase (R06974, EC 6.3.1.21), 2-(Formamido)-N1-(5-phosphoribosyl)acetamidine cyclo-ligase (R04208, EC 6.3.3.1).

Models of primitive thioester and redox coenzymes

Previous work suggested that primitive versions of CoA-coupled and NAD(P)-coupled reactions could have enabled the generation of metabolic networks before the availability of phosphate (*15*, *16*). We explored whether there primitive coenzyme systems could enable purine production by (i) substituting all CoA-coupled reactions with the thiol mercaptopyruvate, and (ii) substituting all NAD(P)H-coupled reactions with hydrogen gas. Note for (i), we were unable to re-compute standard molar free energies because the thioesters formed from mercaptopyruvate were not in the KEGG database. For (ii), we computed standard molar free energies as described previously.

Extending KEGG with additional reactions from ATLAS

We obtained a list of all reactions in the ATLAS database (*30*, *31*). We removed reactions that matched KEGG reactions, resulting in an additional 20,183 reactions. Note that ATLAS also contained an additional 3024 metabolites in the database, increasing the total network to 32,449 reactions and 11,768 reactions.

Seed set

All seed molecules are listed in Table S2. The seed set consisted of all inorganic divalent and monovalent metal species in KEGG, as well as elemental sources of phosphorus, nitrogen, sulfur, oxygen, hydrogen, and carbon. While ammonia, orthophosphate and hydrogen sulfide were the only nitrogen, phosphorus, and sulfur sources provided in the seed set, respectively, several organic compounds were included as seed molecules. Since we did not assume a primitive thioester coupling mechanism and include thiols in the seed set, organic molecules such as pyruvate were necessary to enable expansion (*16*). We thus included organic molecules producible when reacting glyoxylate with pyruvate (*26*), resulting in 19 organic compounds (Table S1). Note that when we assumed that succinate semialdehyde could be producible from succinate via geochemically available reductants (e.g., H_2), we were able to generate networks >4000 compounds with only pyruvate as the organic seed molecule.

Network expansion algorithm

The network expansion algorithm was run as described previously (*15*, *16*). Briefly, seed compounds are allowed to react given the reactions in the network, which then produce product compounds. The product compounds are added to the seed set, and the process is repeated until convergence. The network expansion algorithm was implemented using the `networkExpansionPy` python package (https://github.com/jgoldford/networkExpansionPy).

Identification of purine-containing molecules

Purine-containing compounds were identified using the SUBCOMP search function hosted by the KEGG database (https://www.genome.jp/tools/subcomp/). The query structure was purine (i.e., KEGG compound C15587), the database was "COMPOUND", and the search mode was "SUBstructure".

Defining metabolic pathways

We defined metabolic pathways using KEGG modules (https://rest.kegg.jp/list/module). The ordering of reactions and reaction definitions for each module were determined from the "Reaction" field and "Diagram" field in each KEGG module entry. Data was retrieved using the TogoWS REST service via Biopython (https://biopython.org/docs/latest/api/Bio.TogoWS.html). Because the complexity and non-linearity of some modules make it hard to interpret which reactions are strictly necessary, we limit our analysis to linear pathways (Table S13). A full list of the KEGG modules included for analysis can be found in Table S14. Pathways containing less than 2 reactions are not included. For the linear pathways, we used the reaction to module step rules to compute when each step in the pathway was feasible during the expansion. We then computed the spearman rank correlation between iteration number and pathway step (indexed as 0 to *N*, where *N* is the total number of steps in a pathway). A spearman rank correlation of 1 indicated the pathway emerged sequentially in the forward direction, while a spearman rank correlation of -1 indicated the pathway emerged sequentially in the reverse direction.

Molecular complexity and other ancient features

To compute the molecular complexity for each compound in the expansion scope with a defined molecular formula (*n*=3588), we used RDKit (https://www.rdkit.org/) to compute the Bertz complexity (*72*) using the `GraphDescriptors` module, and used default parameters (Table S8). We computed the average state of reduction per carbon (*2*) (see Fig. S4, Table S9) on all organic molecules with only carbon, oxygen and hydrogen atoms (*n*=1178).

We computed the proportion of microbial taxa using each compound in the expansion by using the KEGG REST API to download KEGG Modules used in each prokaryotic genome (*n*=6416), as well as all KEGG compounds used in each KEGG module. We took 1,312 KEGG compounds that are intermediates in 377 KEGG modules, and computed the proportion of prokaryotic genomes that use each compound as an intermediate in at-least one KEGG module. For all compounds produced and a specific expansion iteration, we computed the average proportion of genomes using each compound (Fig. 2B).

Calculating Reaction-Fold Associations

KEGG orthologous (KO) groups, 8030 in total, were clustered at 80% sequence identity by CD-HIT (*73*) using the default settings and a word size of 5 characters. Using the curated hidden Markov models (HMMs) from the Evolutionary Classification of Domains (ECOD) database (*74*, *75*) (ECOD version develop279), protein families were mapped onto the representative KEGG genes of each orthologous group by the program hmmsearch (*76*). HMM searches used the default settings and the search space (the -Z flag) was set to 106,052,079 sequences. Only domains with an independent E-value of less than $1x10^{-4}$ and an HMM coverage of greater than 70% were considered potential hits. Domains were assigned to regions of a gene in order of increasing E-value using the “envelope coordinates” of the associated hit. If two hits overlap by more than 30% of either hit, the hit with the higher E-value is discarded. Finally, reconciled hits associated with a gene were mapped to their corresponding ECOD X-groups when such mappings are unambiguous. An association between an X-group and KO group was identified if at least 20% of the genes contained that X-group domain.

We next constructed a mapping between reactions and KO groups. We used the KEGG REST API to download the associations between KEGG reactions and Enzyme Commission (EC) numbers, and the EC numbers and KEGG orthologous groups. A KEGG orthologous group was associated with a KEGG reaction if they were associated with at least one shared E.C. number. Finally, X-group and reactions associations were obtained via their mutual associations with KO groups.

Computing structural repetitiveness of protein folds

Protein fold repetitiveness for each representative ECOD domain (70% sequence identity cutoff, ECOD version develop279) was calculated using CE-Symm 2.0 (*56*) with default settings. All classifications other than C1 (i.e., no internal symmetry) were defined as repetitive. Repetitiveness scores were taken to be the fraction of repetitive representative domains within a fold, defined here as an ECOD X-group (Table S11). The average repetitiveness score for a reaction was calculated by averaging the repetitiveness scores of all associated folds (Table S11). The average fold repetitiveness per expansion iteration (Fig. 3C) was computed by averaging across all reactions that emerged at each iteration step.

Statistical analysis

Pearson correlation, spearman correlations and nonparametric Mann Whitney *U* tests were all performed using the `pearsonr`, `spearmanr` and `mannwhitneyu` functions from the `scipy.stats` python module, version 1.5.4 in python 3.6. For Monte Carlo permutation tests performed in Fig. 3B, we used the `random` function in the python native library `sample,` and randomly sampled 10^3 random combinations of 8 molecules from the expansion scope (but not the seed set in Table S6), and repeated the expansion. We estimated a *p*-value by computing the proportion of randomly chosen seed sets that led to an expansion with fewer expansion iterations than the original expansion shown in Fig. 2B (blue line).